Author Accepted Manuscript: Accepted for publication in Journal of the Audio Engineering Society on 25 Oct 18

Capturing 360° Audio using an Equal Segment Microphone Array (ESMA)

Hyunkook Lee, *AES Fellow*
(h.lee@hud.ac.uk)

Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield, HD1 3DH, United Kingdom.

The equal segment microphone array (ESMA) is a multichannel microphone technique that attempts to capture a sound field in 360° without any overlap between the stereophonic recording angle of each pair of adjacent microphones. This study investigated into the optimal microphone spacing for a quadraphonic ESMA using cardioid microphones. Recordings of a speech source were made using the ESMAs with four different microphone spacings of 0cm, 24cm, 30cm and 50cm, based on different psychoacoustic models for microphone array design. Multichannel and binaural stimuli were created with the reproduced sound field rotated with 45° intervals. Listening tests were conducted to examine the accuracy of phantom image localization for each microphone spacing, in both loudspeaker and binaural headphone reproductions. The results generally indicated that the 50cm spacing, which was derived from an interchannel time and level trade-off model that is perceptually optimized for 90° loudspeaker base angle, produced more accurate localization results than the 25cm and 30cm ones, which were based on conventional models derived from the standard 60° loudspeaker setup. The 0cm spacing produced the worst accuracy with the most frequent bimodal distributions of responses between the front and back regions. Analyzes of the interaural time and level differences of the binaural stimuli supported the subjective results. In addition, two approaches for adding the vertical dimension to the ESMA (ESMA-3D) were devised. Findings from this study are considered to be useful for acoustic recording for virtual reality applications as well as for multichannel surround sound.

0 INTRODUCTION

Microphone array techniques for surround sound recording can be broadly classified into two groups: those that attempt to produce the continuous phantom imaging around 360° in the horizontal plane and those that treat the front and rear channels separately (i.e., source imaging in the front and environmental imaging in the rear) [1]. In conventional surround sound productions for home cinema settings, the front and rear separation approach tends to be used more widely due to its flexibility to control the amount of ambience feeding the rear channels. However, with the recent development of virtual reality (VR) technologies that allow the user to view visual images in 360°, the need for recording audio in 360° arises.

Currently, the most popular method for capturing 360° audio for VR is arguably the first order Ambisonics (FOA). FOA microphone systems are typically compact in size, thus convenient for location recording, and offers a stable localization characteristic due to its coincident microphone arrangement [1]. Furthermore, the FOA allows one to flexibly rotate the initially captured sound field in post-production. However, it is known that the FOA has limitations in terms of perceived spaciousness and the size of sweet spot in loudspeaker reproduction due to the high level of interchannel correlation [2]. Higher order Ambisonics (HOA) offers a higher spatial resolution than the FOA and therefore can overcome the limitations of the FOA to some extent, although it is more costly and requires a larger number of channels. A HOA recording can be made using a spherical microphone array (e.g., mh Acoustics

Eigenmike). A system that supports a higher order typically requires a larger number of microphones to be used on the sphere. A review of currently available Ambisonics microphone systems can be found in [3].

On the other hand, a near-coincident microphone array, which incorporates directional microphones that are spaced and angled outwards, can provide a greater balance between spaciousness and localizability than a pure coincident array. This is due to the fact that it relies on both interchannel time difference (ICTD) and interchannel level difference (ICLD) for phantom imaging [4]. The so-called 'equal segment microphone arrays (ESMAs)', originally proposed by Williams [4,5], are a group of multichannel near-coincident arrays that attempt to produce a continuous 360° imaging in surround reproduction. The ESMAs follow the 'critical linking' concept [5], which assumes that a continuous 360° imaging can be achieved when the stereophonic recording angle (SRA)[1] of each stereophonic segment is connected without overlap. There are three requirements to configure and use an ESMA: (i) all two-channel stereophonic segments of the array must have an equal subtended angle between microphones, (ii) the subtended angles must be the same as the SRA for each segment, and (iii) the loudspeaker array for reproduction must have the same angular arrangement as the microphone array. For example, as illustrated in Fig. 1, a four-channel (quadraphonic) ESMA is configured to produce the SRA of 90° using four unidirectional microphones with the subtended angle of 90° for each stereophonic segment; each of the microphone signals is discretely routed to each loudspeaker in a quadraphonic setup. Although the ESMA was originally proposed as a recording technique for multichannel loudspeaker reproduction [4,5], it is proposed here that the ESMA would also be suitable for binaural headphone reproduction with head tracking for 360° audio applications. This can be achieved by convolving the ESMA signals with head-related transfer functions (HRTFs) for the corresponding loudspeaker positions, which are dynamically updated according to the angle of head rotation.

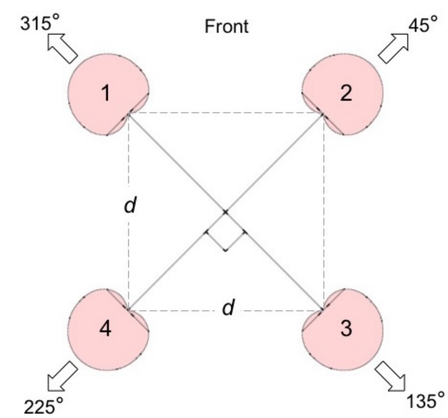

Fig. 1. Top view of a quadraphonic equal segment microphone array (ESMA) using cardioid microphones. The microphone spacing (*d*) is determined to produce the stereophonic recording angle of 90°.

The current study[2] aims to (i) determine the appropriate microphone spacing for a quadraphonic ESMA using cardioid microphones to achieve an SRA of 90° and (ii) examine the localization characteristics of the ESMA in loudspeaker and binaural headphone reproductions with sound field rotations. The spacing and subtended angle between microphones for a microphone array with a specific SRA are determined based on a psychoacoustic ICTD and ICLD trade-off relationship required for a full phantom image shift, as discussed in details in Section 1. In case of the ESMA, the subtended angle between microphones is predetermined according to the number of channels involved (e.g., 90° for four channels) as mentioned above, thus making the microphone spacing the sole factor to determine the SRA. For example, if a correct microphone spacing is applied to the quadraphonic ESMA, then a sound source located at ±45° will be localized at ±45° in a quadraphonic reproduction with 90° base angle for each stereophonic segment. There exist several different ICTD and ICLD trade-off models for estimating the SRA [8,9,10], and it is of interest of this study to discover which model produces the most accurate result. Conventional models [8,9] have been derived from experimental data obtained using the standard 60° loudspeaker setup. However, each stereophonic segment in the quadraphonic reproduction for the ESMA has the base angle of 90°. Therefore, the validity of applying such models to the design of the ESMA is questioned here. From this, the present study evaluates the imaging accuracies of the quadraphonic cardioid ESMAs with four different microphone spacings

[1] The SRA refers to the horizontal span of the sound field in front of the microphone array that will be reproduced in full width between two loudspeakers [6].
[2] Preliminary results from this work were presented at the AES International Conference on Audio for Virtual and Augmented Reality in 2016 [7].

based on different models: (i) 24cm based on both the Williams curves [8] and Image Assistant [9] models, both of which are based on data obtained using the 60° loudspeaker setup, (ii) 30cm based on the Microphone Array Recording and Reproduction Simulator (MARRS) model [10], which is also originally derived from the 60° setup, (iii) 50cm based on the MARRS model that is perceptually optimized for the 90° setup, and (iv) 0cm as in the so-called 'in-phase' decoding of the FOA B-format signals [2], which is equivalent to using four cardioids arranged in the quadraphonic setup.

The rest of the paper is organized as follows. Section 1 discusses the psychoacoustic models used to calculate different microphone spacings for the quadraphonic cardioid ESMAs tested. Section 2 describes methods used for two listening experiments conducted in loudspeaker and binaural headphone reproductions. Results obtained from the experiments are statistically analyzed in Section 3, followed by the discussions of the results in Section 4. Section 4 also analyzes interaural time and level difference cues in the binaural stimuli and discusses possible ways to extend the ESMA for three dimensional sound recording. Finally, Section 5 concludes the paper.

1 PSYCHOACOUSTIC MODELS

This section describes three different ICTD and ILD trade-off models that were used to derive the microphone spacings tested in this study.

1.1 Williams Curves

Williams [5] recommends the microphone spacing of 24cm for the quadraphonic cardioid ESMA. This is estimated based on the so-called 'Williams curves' [8], which are a collection of curves that indicate possible combinations of microphone spacings and subtended angles to achieve specific SRAs. They are based on an ICTD and ICLD trade-off relationship derived from the polynomial interpolations of ICTD and ICLD values required for 10°, 20° and 30° image shifts that were obtained from a listening test in the standard 60° loudspeaker setup. Williams [8] claims that the SRA is virtually independent of the loudspeaker base angle, suggesting that the same ICTD and ICLD trade-off model obtained for the 60° loudspeaker setup can also be applied to the 90° setup. From this, he proposes that 24cm between each microphone in the quadraphonic cardioid ESMA can produce the desired SRA of 90° for each stereophonic segment. Note that the ICTD and ICLD produced by a near-coincident microphone configuration vary slightly depending on the distance between sound source and microphone array, and so does the SRA of the array. However, it is not stated in [8] what source-array distance the Williams curves were based on.

1.2 Image Assistant

In contrast with the Williams's curves, the psychoacoustic model used in the 'Image Assistant' tool [9] assumes a linear trade-off between ICTD and ICLD within the 75% image shift region (e.g., 0 to 22.5° for the 60° loudspeaker setup). It also allows the user to choose a specific source-array distance for the SRA estimation. The amount of total image shift within this region is estimated by simply adding the image shifts that individually result from ICTD and ICLD (13%/0.1ms and 7.5%/dB, respectively), which is a method proposed by Theile [11]. Outside the linear region, where the image shift pattern tends to become logarithmic for both ICTD and ICLD, an approximate function is applied to derive a non-linear ICTD and ICLD trade-off relationship [12]. The tool suggests that at 2m distance between the source and the centre of the array, which was used in the experiment of the present study, 24cm is the correct microphone spacing to produce the required SRA of 90°. The ICTD and ICLD shift factors used in the Image Assistant were obtained for the standard 60° loudspeaker setup. However, as in William's assumption that the SRA is conserved regardless of the loudspeaker base angle, Theile [13] also claims the same ICTD and ICLD image shift factors can be used for an arbitrary loudspeaker base angle, which is here referred to as the constant relative shift theory. Based on this, the microphone spacing of 24cm is assumed to be still valid for the loudspeaker base angle of 90° in the quadraphonic reproduction setup.

1.3 MARRS

The 30cm and 50cm spacings are based on SRA estimations using the present author's microphone array simulation tool 'MARRS (Microphone Array Recording and Reproduction Simulator)' [10]. The psychoacoustic model used for MARRS relies on an ICTD and ICLD trade-off model derived from region-adaptive ICTD and ICLD image shift factors for the 60° loudspeaker setup presented in Table 1; they were defined based on subjective localisation test data obtained using natural sound sources [14]. If Theile's constant relative shift theory described above is applied here (i.e., using the data obtained for the 60° loudspeaker setup for the 90° setup), the correct spacing for each segment of the quadraphonic cardioid ESMA to achieve the 90° SRA at 2m source-array distance is 30cm.

Table 1. ICTD and ICLD shift factors for the 60° and 90° loudspeaker setups suggested by the MARRS psychoacoustic model [10].

Speaker base angle	Image shift region	Shift factor	
		ICTD	ICLD
60°	0–66.7%	13.3%/0.1ms	7.8%/dB
	66.7%–100%	6.7%/0.1ms	3.9%/dB
90°	0–66.7%	8.86%/0.1ms	6%/dB
	66.7%–100%	4.43%/0.1ms	3%/dB

However, the author's previous research on amplitude panning [15] suggests that ICLD shift factors must vary depending on the loudspeaker base angle in order to achieve an accurate phantom image localization; a larger base angle requires a larger ICLD for a given proportion of image shift. An informal listening test confirmed that this was also the case with ICTD. Therefore, the MARRS model [10] scales the original ICTD and ICLD shift factors depending on the loudspeaker base angle. For example, for the 90° loudspeaker setup, the original ICLD shift factor is scaled by 0.77, which is the ratio of the interaural level difference (ILD) above 1 kHz produced at 30° (the loudspeaker azimuth in the original 60° setup, which serves as the reference) to that at 45° (the loudspeaker azimuth of the 90° setup). Similarly, the ICTD shift factor is multiplied by the ratio of interaural time differences (ITDs) below 1 kHz between 30° and 45° (0.67). This scaling process results in shift factors optimized for the 90° loudspeaker setup, which are presented in Table 1. Based on these, the correct spacing between adjacent microphones for the quadraphonic cardioid ESMA is estimated to be 50cm. Note that this spacing is calculated for the source-array distance of 2m. However, the difference for a larger distance in a practical recording situation is very small, e.g., 50.4cm spacing for 5m source distance. Readers who are interested in the detailed algorithm used in MARRS is referred to the open-access Matlab source code package[3]. MARRS is also available as a free mobile app from the Apple and Google app stores.

2 EXPERIMENTAL DESIGN

Two subjective experiments were carried out. Experiment 1 evaluated the localization accuracies of the four microphone arrays with different spacings in a quadraphonic loudspeaker reproduction. Experiment 2 repeated the same tests over headphones using binaurally synthesised stimuli of the ESMAs. Various degrees of head rotations were simulated by rotating the reproduced sound field by the corresponding degrees with the listeners kept facing forwards. This method was opted over real head-rotations since it allowed an efficient randomisation and accurate implementation of target angle condition for each trial. Furthermore, the head-static listening with sound field rotation is a practical scenario, e.g., watching 360° video on a monitor screen rather than using a head-mount display. However, results from this study would require verification in a practical virtual reality scenario with head tracking in the future.

2.1 Physical Setup

The experiments were conducted in the ITU-R BS.1116-compliant listening room of the Applied Psychoacoustics Laboratory at the University of Huddersfield (6.2 x 5.6 x 3.8m; RT = 0.25s; NR = 12). The room was used for both stimuli creation and listening tests. Eight Genelec 8040A loudspeakers were arranged in a circle as shown in Fig.2. The loudspeakers were positioned at the azimuth angles of 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315° clockwise. The distance between the center of the circle (the listening position) and each loudspeaker was 2m. In the listening tests, the loudspeaker setup was hidden to the listeners by using acoustically transparent curtains.

[3] https://github.com/APL-Huddersfield/MARRS

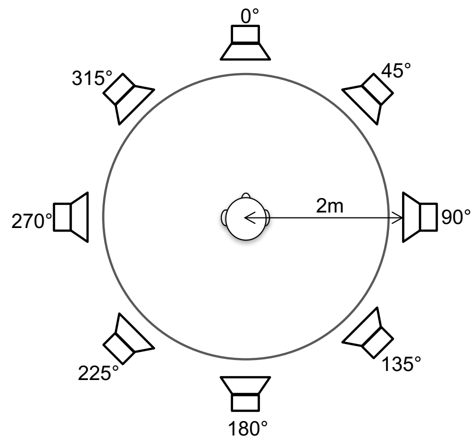

Fig. 2. Loudspeaker setup used for room impulse response measurements and Experiment 1.

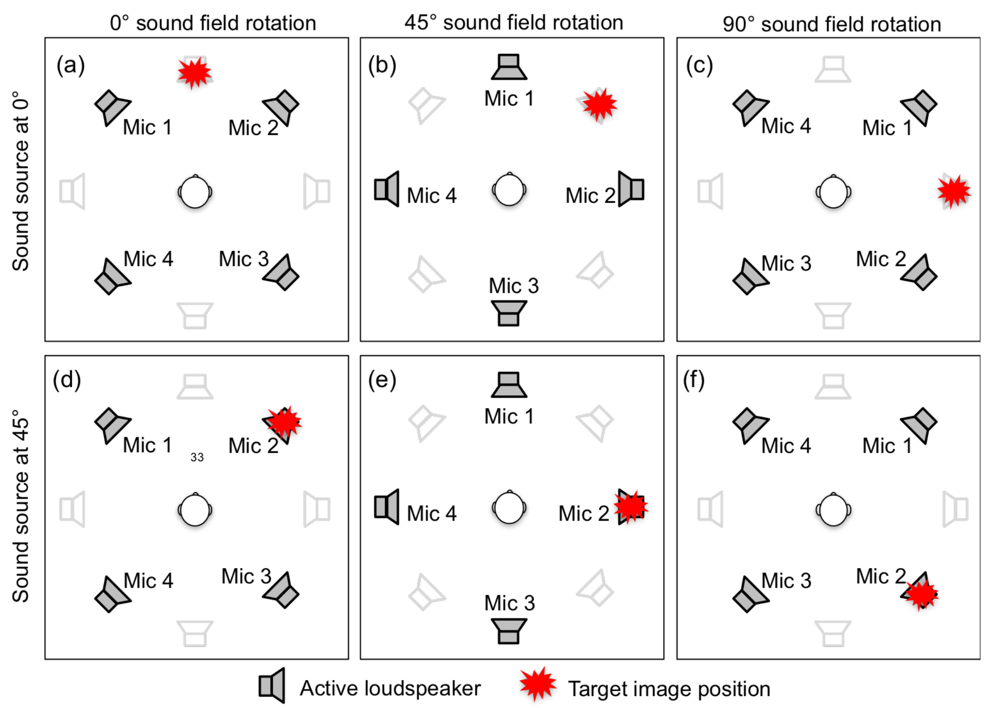

Fig. 3. Examples of sound field rotation applied to stimuli created for sound sources at 0° and 45°; each sound field rotation simulating the equivalent head rotation.

2.2 Stimuli Creation

2.2.1 Room impulse response measurement

In order to create test stimuli, four-channel room impulse responses (RIRs) were first acquired in the listening room for each of the four microphone arrays individually, using the exponential sine sweep method [16]. The microphones used for the ESMAs with 24cm, 30cm and 50cm were Neumann KM184 cardioid microphones, which were pointing towards 45°, 135°, 225° and 315°. In addition to the ESMAs, a Soundfield SPS422b FOA microphone system was used to capture B-format RIRs, which were decoded using the in-phase decoding method [2] as mentioned earlier. This produced the quadraphonic cardioid polar responses of four virtual microphones that were coincidentally arranged and pointing towards 45°, 135°, 225° and 315°.

Sound sources used for the RIR measurements were the loudspeakers placed at 0° and 45°. They were selected for the following reasons. Firstly, the 45° position was to investigate whether the arrays could achieve the goal of the 90° SRA for each stereophonic segment. If the goal were indeed achieved, then the phantom image for the source should be localized at 45° in reproduction. The 0° position was selected for examining how accurately a centrally panned phantom image can be localized at the desired position for a given sound field rotation.

2.2.2 Stimuli for experiment 1

For the loudspeaker listening test, four-channel stimuli for each source position were created by convolving the RIRs captured using the microphones with an anechoically recorded male speech signal taken from [17]. Prior to the convolution, all reflection components of the RIRs (i.e., beyond 2.5ms after the direct sound) were removed using a half Hann window. This was to avoid excessive room reflections to be heard when the stimuli were reproduced in the same room where the RIRs were captured. However, it should be acknowledged that in practical situations the recording and reproduction environments are usually different and their acoustic characteristics would interact.

Table 2. Target image position for each sound field rotation for each source position

Source position	Sound field rotation	Equivalent head rotation	Target image position
0°	0°	0°	0°
	45°	-45°	45°
	90°	-90°	90°
	135°	-135°	135°
	180°	-180°	180°
	225°	-225°	225°
	270°	-270°	270°
	315°	-315°	315°
45°	0°	0°	45°
	45°	-45°	90°
	90°	-90°	135°
	135°	-135°	180°
	180°	-180°	225°
	225°	-225°	270°
	270°	-270°	315°
	315°	-315°	0°

Sound field rotations from 0° to 315° were applied to the original four-channel stimuli with 45° intervals. This was done by offsetting the azimuth of the loudspeaker for each of the four signals by 45° for every 45° rotation. For instance, as illustrated in Fig. 3 (c) and (f), the signals of microphones 1, 2, 3 and 4 shown in Fig. 1 were presented from the loudspeakers at 45°, 135°, 225° and 315°, respectively, for a 90° sound field rotation. In this case, the target perceived positions for the sound sources at 0° and 45° were 90° and 135°, respectively. Table 2 presents the target image position for each sound field rotation and its equivalent head rotation for each source position.

In addition, eight real source stimuli were created by routing the speech signal to each of the eight loudspeakers individually. These served as reference conditions to compare the localization behaviors of the phantom source stimuli against.

2.2.3 Stimuli for Experiment 2

For the binaural listening test, the same speech signal used in Experiment 1 was convolved with the RIRs captured using the microphone arrays. In contrast with the loudspeaker listening test, full RIRs including room reflections were used to auralise the listening room condition. The resulting signals were then convolved with anechoic head-related impulse responses (HRIRs) captured using a Neumann KU100 dummy head, which were taken from the 'SADIE' database [18]. Head rotations were simulated by applying HRIRs corresponding to the target position associated with each rotation angle. Additionally, reference binaural stimuli for a real source were created by recording the anechoic speech reproduced from each of the eight loudspeakers in the listening room using a Neumann KU100 dummy head placed at the listening position.

2.3 Subjects

Nine critical listeners participated in both experiments, in which they tested each stimulus condition twice in a randomized order for each experiment; a total of 18 localization responses were obtained for each test condition. They comprised staff researchers, postgraduate research students and final year undergraduate students of the Applied Psychoacoustics Lab at the University of Huddersfield, with their ages ranging from 21 and 38. All of them reported normal hearing and had extensive experience in conducting sound localization tasks in formal listening tests. All subjects completed the loudspeaker test (Experiment 1), at least one week after which they sat the binaural test (Experiment 2). they did not know the nature of the test stimuli until they completed both experiments.

2.4 Test Procedure

2.4.1 Experiment 1

The subject was seated at the center of the loudspeaker circle, and the chair was adjusted so that their ear height matched the height of the loudspeaker's acoustic center (1.35m from the floor). The subjects were instructed to face the front and not to move their heads during the test, while eye movement was encouraged. A small headrest was placed at the back of the subject's head to reduce movement, which was visually monitored by the experimenter during the test. The subject's task was to mark down the apparent location of perceived image for each stimulus on a horizontal circle provided on a graphical user interface (GUI) written using Max 7. The angular resolution in the response was 1°. Small markers were indicated on the circle from 0° with 22.5° intervals. Markers with the same intervals were also placed on the acoustic curtains to help the subject correctly map the perceived image position on to the circle. Prior to the actual test, the subjects were given familiarization trials comprising the real source stimuli for the eight loudspeaker positions, which were considered to have the highest localization accuracy amongst all stimuli.

The playback levels of all stimuli were calibrated to 70dB LAeq at the listening position. Each trial in the test contained a single stimulus and the subjects could listen to it repeatedly until they judged its perceived position. All stimuli were presetented in a randomised order. For the sound-field-rotated stimuli, one of the mirrored target image positions (e.g., 315° or 45°) was randomly selected for each listener for each microphone array condition. This was to minimise psychological order effects as well as to avoid a potential listening fatigue that might occur when the sound is presented only from the left or right-hand side. Every subject judged each test condition twice in a randomized order.

2.4.2 Experiment 2

The listening test was conducted in the same room as Experiment 1. The test procedure was identical to that of Experiment 1, apart from the following. The headphones used for the test were Sennheiser HD650. To equalize them, their impulse responses were measured five times using the KU100 dummy head, with them re-seated on the head each time. The average responses were then inverse filtered using a regularization method by Kirkeby et al. [19]. Prior to the actual test the subjects were presented with familiarization trials comprising the binaural recordings of the real sources for the eight loudspeaker positions. The loudness unit level of all binaural stimuli was calibrated at -18 LUFS and the headphone playback level was determined by the present author to match the perceived loudness of the loudspeaker playback from Experiment 1 (70dB LAeq). No head tracking was used for rendering different image positions in binaural reproduction; the sound field was rotated instead as described in Section 2.2.3.

3 RESULTS

As mentioned earlier, the stimuli with the mirrored target image positions were randomly selected for each listener in the listening tests. For the purposes of the statistical analysis and data plotting, the perceived angles for the stimuli with the target angles in the left-hand side of the circle were converted into the corresponding angles in the right-hand side (e.g., 315° to 45°, 270° to 90°). For the continuity of data in the analysis, any responses for the 0° target angle that were given in the left-hand side of the circle were converted into negative values (e.g., 355° to -5°), whereas those for the 180° target angle in the left side were unchanged.

Shapiro-Wilk and Levene's tests were first performed to examine the normality and variance of the data collected. The results suggested that the data were not suitable for parametric statistical testing. Therefore, the non-parametric Wilcoxon signed-rank test was conducted to examine if there was a significant difference between the target and perceived image positions for each test condition, except for those that had a significant bimodal distribution. The significance of bimodality was examined using the Hartigan's dip test [20].

3.1 Phantom Source Localization in Loudspeaker Reproduction

Fig. 4 shows the scatter plots of the data for the phantom source conditions (i.e., microphone array recordings) from Experiment 1. Table 3 presents the summary of the statistical analyzes.

3.1.1 Sound source at 0°

The results for the 0° source position are first presented. From the scatterplots in Fig. 4, it appears that all microphone spacings produced a relatively accurate localization when the target angle was 0°; there is no front-back confusion. For the 45° target angle (45° simulated head rotation), the 0cm condition had the median perceived angle (MED) of 24°, which was significantly smaller than the target (p = 0.027), whereas the differences of the 50cm, 30cm and 24cm spacings to the target was not significant (p > 0.05). Looking at the 90° target angle (90° simulated head rotation), the responses for the 0° source appear to have wide spreads in general. The 50cm spacing had a significant bimodal distribution (p = 0.022). The MEDs for the 30cm and 24cm were considerably smaller than the target angle (67°–68°). The 0cm spacing had the largest deviation from the target angle amongst all spacings (MED = 45°, p = 0.015). For both the 135° and 180° target angles, the MEDs for all spacings did not have a significant difference from the target angles (p > 0.05). However, the responses for the 135° target angle tended to be widely spread between the front and rear regions.

Table 3. Summary of the results for phantom source localization in loudspeaker reproduction (Experiment 1): Median perceived angles for each experimental condition. Conditions with a significant difference from the target position (Wilcoxon signed rank test): * p < .05; ** p < .01. Conditions with a significant bimodal distribution (Hartigan's dip test): ^ p < .05; ^^ p < .01.

		Target azimuth after sound field rotation (degree)				
Source angle (degree)	Mic spacing (cm)	0	45	90	135	180
0	50	0	41	^	135	180
	30	0	40	67*	134	180
	24	0	34	68	135	180
	0	0	24**	45*	134	180
45	50	0	45	90	135	180
	30	0	44*	90	135	^
	24	0	39**	90	135	180
	0	0	30**	^^	152	^

3.1.2 Sound source at 45°

For the 0° target angle (315° sound field rotation), all conditions had no significant difference between the perceived and target angles (p > 0.05). For the 45° target angle (no sound field rotation), the MED was closer to the target angle in the order of 50cm (45°), 30cm (44°), 24cm (39°) and 0cm (30°). Apart from the 50cm spacing, the MEDs were all found to deviate significantly from the target (p = 0.047 for 30cm, p = 0.000 for 24cm and 0cm). For the 90° target angle, the 50cm, 30cm and 24cm spacings did not have a significant difference between the perceived and target angles (MED = 90°, p > 0.05), whereas the 0cm produced a significant bimodal distribution between around 45° and 135° (p = 0.002). Looking at the target angle of 135°, the MEDs for the 50cm, 30cm and 24cm were the same as the target, whereas that for the 0cm (152°) was noticeably closer to the median plane, although this was not statistically significant (p > 0.05). For the 180° target angle, 50cm and 24cm were found to produce an accurate result (MED = 180°, p > 0.05), whereas responses for 30cm and 0cm had a significant bimodality (p = 0.036 and 0.01, respectively).

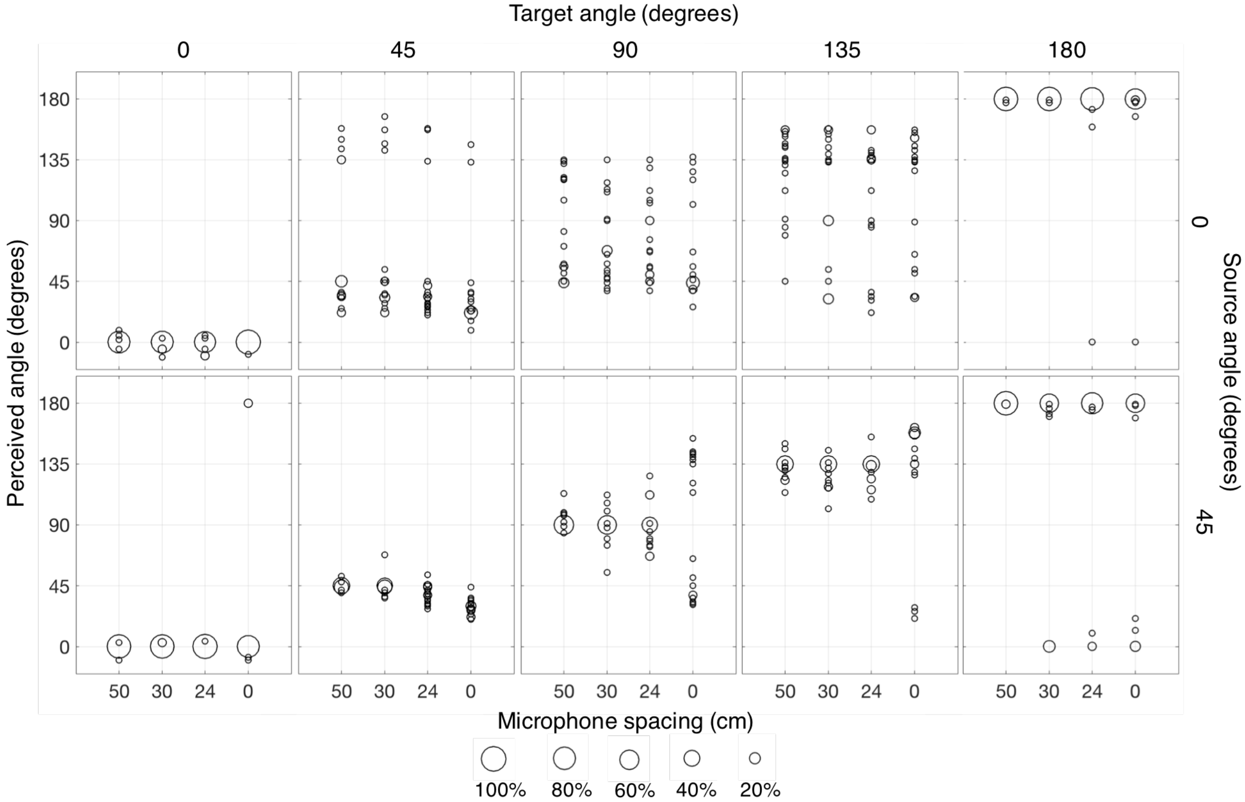

Fig. 4. Bubble plots of the data obtained from the loudspeaker localization test (Experiment 1). The diameter of each circle represents the percentage of responses for each condition.

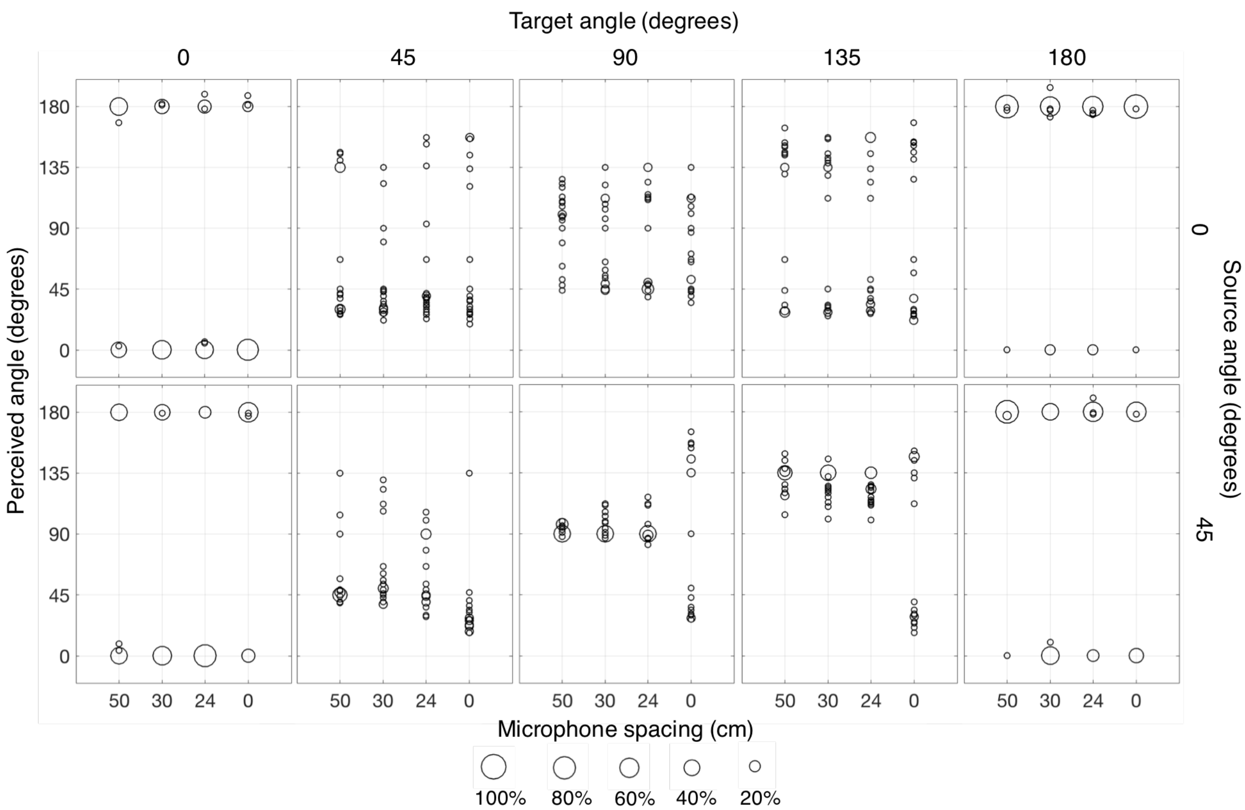

Fig. 5. Bubble plots of the data obtained from the binaural localization test (Experiment 2). The diameter of each circle represents the percentage of responses for each condition.

3.2 Phantom Source Localization in Binaural Reproduction

The scatter plots of the data obtained for the phantom source conditions from Experiment 1 are presented in Fig. 5. Table 4 summarizes the results from the statistical analyzes. From Fig. 5, it is generally observed that the responses from the binaural test were more widely spread compared to those from the loudspeaker test (Fig. 4). Table also indicates that the binaural test had more conditions with a significant bimodal distribution.

Table 4. Summary of the results for phantom source localization in binaural reproduction (Experiment 1): Median perceived angles for each experimental condition. Conditions with a significant difference from the target position (Wilcoxon signed rank test): * $p <$ .05; ** $p <$ .01. Conditions with a significant bimodal distribution (Hartigan's dip test): ^ $p <$ .05; ^^ $p <$ .01.

		Target azimuth after sound field rotation (degree)				
Source angle (degree)	Mic spacing (cm)	0	45	90	135	180
0	50	^^	42	100	^^	180
	30	^^	35	62	^^	180
	24	^^	39	^	^	180*
	0	^^	39	69	^	180
45	50	^^	47	90	135	180
	30	^^	50*	90*	129**	^^
	24	^	47	90	^	^
	0	^^	27*	^^	^^	^^

3.2.1 Sound source at 0°

Looking at the results for the 0° source first, the responses for the 0° target were significantly bimodal for all of the spaced array conditions ($p < 0.01$). The responses were mainly given to either 0° or 180°, exhibiting strong tendencies of front-to-back confusion. For the target angle of 45°, none of the spacings produced a significant difference between the perceived and target angles, although 50cm had a MED that is closest to the target. For the 90° target angle, again the 50cm spacing produced the most accurate result. The MEDs for 30cm and 0cm (62° and 69°, respectively) were considerably narrower than the target, while responses for 24cm were significantly bimodal ($p < 0.05$). All conditions for the target angle of 135° were found to have a significant bimodal distribution between around 45° and 135° ($p < 0.05$ for 50cm and 30cm, $p < 0.01$ for 24cm and 0cm) . For the 180° target angle, only the 30cm condition was found to be significantly different from the target ($p < 0.05$).

3.2.2 Sound source at 45°

For the 45° source position, the responses for the target angle of 0° were found to be significantly bimodal regardless of the microphone spacing (i.e., front-to-back confusion). For the 45° target angle, the 50cm and 24cm spacings both produced the MED of 47°, which was not significantly different from the target ($p > 0.05$). However, the 30cm and 0cm had significant differences between the target and perceived angles (MEDs = 50° and 27°, respectively, $p < 0.05$). The results for the 90° target angle show that the 50cm, 30cm and 24cm all had the median perceived angles of 90°, whereas the 0cm condition had a significant bimodal distribution ($p < 0.01$) between around 45° and 135°. For the 135° target angle, 50cm was the only spacing that produced an accurate result (MED = 135°, $p > 0.05$). The MED for 30cm (129°) was significantly different from the target ($p = 0.007$), while 24cm and 0cm had a significant bimodal distribution ($p = 0.04$ for 24cm and 0.000 for 0cm). Lastly, for the target angle was 180°, the 50cm spacing produced an accurate result (MED = 180°, $p > 0.05$), whereas the other spacings all had a significant bimodality.

3.3 Real Source Localization in Loudspeaker and Binaural Reproductions

Fig. 6 presents the responses given to the real source stimuli (i.e., single loudspeaker conditions) in both loudspeaker and binaural experiments. Wilcoxon tests suggest that, for the loudspeaker results, there was no significant difference between the perceived and target angles for all stimuli ($p > 0.05$). For the binaural conditions, on the other hand, it was found that the responses for the 0° and 180° sources were significantly bimodal, exhibiting front-back confusion. Furthermore, the 45° source (MED = 52°) was found to be perceived at a significantly wider position than the target ($p < 0.01$).

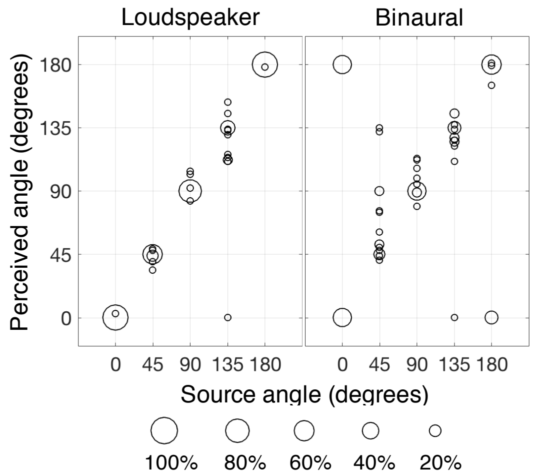

Fig. 6. Bubble plots of the data obtained for single sources from the loudspeaker and binaural tests. The diameter of each circle represents the percentage of responses for each condition.

4 DISCUSSIONS

This section discusses various aspects of the subjective results described above. The measurements of interaural time and level differences are provided to explain the subjective results. A higher order and 3D versions of ESMA are also introduced.

4.1 Microphone Spacing

In general, amongst all of the microphone spacings tested, 50cm produced the best results in terms of phantom image localization accuracy. In the loudspeaker presentation, for all target angle conditions apart from 90°, the 50cm spacing had no significant difference between the target and median perceived angles (MEDs) as evident in Table 2. This seems to validate the ICTD-ICLD trade-off model of the MARRS [8] that is optimized for the 90° loudspeaker base angle (Section 1.3). The 45° source angle with no sound field rotation was a particularly important test condition for examining whether the quadraphonic ESMA can achieve the goal of 90° SRA, as discussed in Introduction. The results indicate that the 24cm and 30cm spacings, which are based on conventional psychoacoustic models [6,7], fail to achieve the goal; they produced significantly narrower MEDs than the target angle of 45°. In the binaural presentation, there were generally more bimodal distributions than in the loudspeaker test. However, 50cm had the most conditions that were not significantly different from the target positions. The differences between the loudspeaker and binaural results are further discussed in Section 4.3.

The 0cm spacing demonstrated the worst localization performance, having the largest number of conditions where the MED was significantly narrower than the target angle or the data distribution was significantly bimodal. For example, the MEDs for the stimulus with the source recorded at 45° were only 30° and 27° in the loudspeaker and binaural presentations, respectively. However, it is worth noting that this should not be assumed as the general localization performance of FOA. As mentioned in Section 2.2.1, the current study used the four virtual cardioid microphones derived from the in-phase decoding of B-format signals. This was for direct comparisons against the ESMAs with cardioid microphones. The polar pattern of virtual microphone formed by the basic (or mode-matching) decoder is a supercardioid [21], which has a higher directionality than a cardioid. Therefore, it is expected that the phantom image would be localized closer to the target position of 45° if the basic decoder was used for the FOA recording.

4.2 Source Angle

The responses for the 0° source tended to have larger data spread and more bimodal distributions than the 45° source, especially when sound field rotations were applied. This could be explained as follows. The ICTD and ICLD trade-off models that the different spacings were based on were originally obtained for a loudspeaker pair that was symmetrically arranged in the front. With a sound field rotation, the signals for the 0° source would create a phantom image between the loudspeakers that are asymmetrical to the direction where the head faces (e.g., Fig. 3). Therefore, the original trade-off models would not be applied correctly. More notably with the 90° rotation of the sound field for the 0° source (90° target angle), where the signals were presented dominantly from the loudspeakers at 45° and 135°, the responses were noticeably spread or bimodal between 45° and 135° in both loudspeaker and binaural conditions. The poor localization certainty of a lateral phantom image observed in the current study is in line with past results reported by Theile and Plenge [22] and Martin et al. [23].

From the above discussion, it might be suggested that, in 360° audio applications with sound field rotation or head-tracking, the localization accuracy and precision of a quadraphonic ESMA might be at their best with sources around the edges of the SRA (i.e., ±45°), and become poorer as the source azimuth becomes closer to ±90°.

4.3 Loudspeaker Reproduction vs. Binaural Reproduction

Overall, the loudspeaker and binaural presentations produced similar patterns of phantom image localization, but Wilcoxon tests performed between the loudspeaker and binaural test data suggest that there were a few conditions that had significant differences. Notably, the 0° target angle condition had a significant bimodality in the binaural presentation for both the 0° and 45° source positions, but not in the loudspeaker presentation. Furthermore, the 45° source condition without a sound field rotation (i.e., 45° target angle) produced responses spread between around 45° and 135° in the binaural reproduction (i.e., front-back confusion), whereas it was localized only in the front region in the loudspeaker reproduction. It is interesting that similar tendencies were also observed for the single sources at 0° and 45° (see Fig. 6). It may be suggested that the front-back confusion observed for the 0° and 45° target angle conditions were associated with the binaural synthesis using the non-personalized HRTSs. However, as Wightman and Kistler [24] found, such confusions could happen even with personalized HRTFs when head movement is not allowed. The current experiment did not allow head movement while listening, which might explain the front-back confusion observed. From the above, it is considered that, in practical VR applications with head tracking, such an issue may be resolved even if non-individualized HRTFs are used for the binaural rendering of ESMA, which requires further investigation.

4.4 Analyzes of Interaural Time and Level Differences

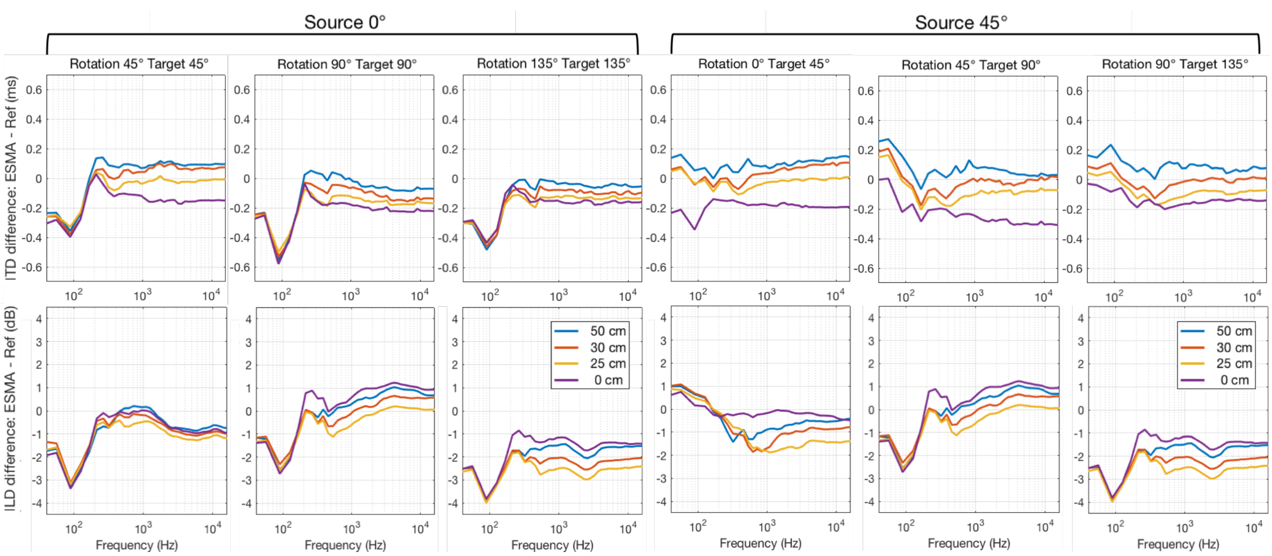

Fig. 7. Difference of ESMA to real source in Interaural time difference (ITD) and interaural level differences (ILD) for each experimental condition; average of results obtained for 50ms overlapping windows for each of the 42 ERB critical bands.

To gain further insights into potential reasons for the subjective results, the ITDs and ILDs of all of the binaural stimuli with off-center target angles (45°, 90° and 135°) were estimated and compared. 0° and 180° were esince at those angles there is no ITD and the ILD exists only at very high frequencies due to ear asymmetry. The binaural model used for the analyzes is described as follows. Each binaural stimulus was first split into 42 frequency bands through a Gammatone 'equivalent rectangular band (ERB)' filter bank [25], which mimics the critical bands of the inner ear. To emulate the breakdown of phase-locking mechanism in the ear signals, half-wave rectification and a first-order low-pass filtering at 1 kHz were applied to each band, as in [26,27]. Time-varying ITD and ILD for each band were computed for 50%-overlapping 50ms frames with the Hanning window. The ITD was defined as the lag of the maximum of the normalized interaural cross-correlation function (i.e., lag ranging between -1 ms and 1 ms). The ILDs were computed as the energy ratio between the left and right signals. The ITDs obtained for all of the frames were averaged for each band, so were the ILDs. The results are presented in Fig. 7 as the ITD and ILD differences of each microphone array stimulus to the real source stimulus with the corresponding target angle (i.e., the single source dummy head recordings). Therefore, the closer the difference is to the 0 reference, the more accurate the ITD or ILD produced by the microphone array is.

Looking at the plots for the 45° source with a 0° rotation (45° target angle), the 50cm spacing produced slightly more ITDs than the dummy head reference across all bands, while it produced slightly lower ILDs constantly above about 200 Hz. It was shown in the subjective results that the 50cm spacing produced an accurate localization for this test condition. Based on the literature [28,29], this subjective result seems to be due to a trade-off between the effects of the ITDs and ILDs on localization. That is, a wider image position due to the ITD being greater than the reference and a narrower image position due to the ILD being smaller than the reference might have been spatially averaged. Especially between about 700 Hz and 4 kHz, where Griesinger [30] claims to be the most important frequency region to determine the perceived position of a broadband phantom image, the average ITD and ILD differences to the reference for this condition are 0.1 ms and -0.75 dB, respectively. This gives the ratio of 0.13 ms/dB, which lies within the range of ITD/ILD trading ratios[4] found in the literature (i.e., 0.04 – 0.2 ms/dB [26]). This suggests that the degree of the positive image shift from the target position by the ITD cue and that of the negative shift by the ILD cue would have been similar, thus resulting in the spatial averaging around the target position. On the other hand, for all the other spacing conditions for the 45° source with a 0° rotation, the 'center of gravity' between the ITD and ILD images (as described in [29]) seem to be at a narrower position than the target. For example, for the 25cm ESMA, the average ITD difference to the reference between 700 Hz and 4 kHz was only -0.02 ms, whereas the average ILD difference was -1.7dB. This would have caused a considerable deviation from the target towards a narrower position mainly due to the ILD cue. It is also interesting to observe that the 0cm condition, which had the worst subjective result, had the opposite trend to the 25cm condition; the average ILD difference was only -0.15 dB, whereas the ITD difference was considerably large (-0.18 ms). A similar trend to the above is generally observed in the other source-rotation conditions.

4.5 Higher Resolution ESMA

The unstable side image localization in head rotations, which was discussed in Section 4.2, could be improved if the SRA resolution is increased. For example, an octagonal (eight-channel) ESMA, which was originally proposed by Williams [5], is considered here.

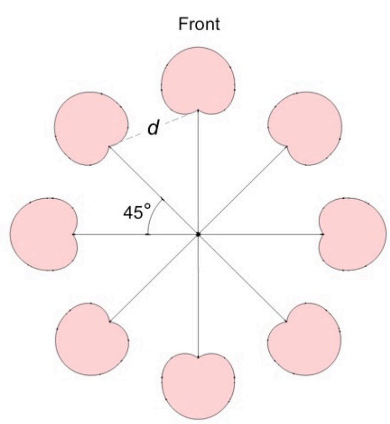

Fig. 8. Octagonal cardioid ESMA. *d* = 82cm according to Williams's ICTD-ICLD trade-off model [8]; 55cm according to the MARRS model [10].

As illustrated in Fig. 8, the microphone array is configured with eight spaced cardioid microphones arranged in an octagon with the 45° subtended angle for each microphone pair. It requires an octagonal loudspeaker layout for reproduction. To achieve the 'critical linking' for each stereophonic segment, the SRA for each pair of adjacent microphones should be made 45°, for which the microphone spacing *d* should be determined. As discussed earlier, different microphone spacings can be suggested depending on which psychoacoustic model for ICLD and ICTD trade-off. If cardioid microphones are used, for example, the necessary spacing is 82cm according to the Williams curves [8], whereas it is 55cm based on the MARRS model [10]. This is because the MARRS scales the ICTD and ICLD trade-off function adaptively depending on the loudspeaker base angle as described in Section 1.3, whereas the Williams curves applies the same model used for the 60° base angle. Further study is required to confirm the localization accuracies of various spacings for the octagonal ESMA.

[4] ITD/ILD trading ratio refers to the equivalence between interaural time and level differences measured in terms of the magnitude of perceived image shift [29].

4.6 ESMA-3D

Two methods of adding the height dimension to the quadraphonic ESMA for 3D sound reproduction (namely, ESMA-3D) are proposed in this section. The underlying design concept for the ESMA-3D is to use horizontally spaced pairs of vertically coincident microphones. The rationale for the choice of the vertically coincident configuration is as follows. Firstly, in terms of vertical source localization, Wallis and Lee [31] showed that a vertical ICTD is an unstable cue for vertical stereophonic panning due to the lack of the precedence effect in the vertical plane. On the other hand, a vertical ICLD was found to have some control over the perceived vertical image position, although its perceptual resolution and consistency were not high [32,33]. Furthermore, Lee and Gribben [34] found that vertical spacing between main and height microphones of a main microphone array had no significant effect on the perceived spatial impression. A vertical coincident design also has an advantage in 3D-to-2D downmixing in that there is no comb-filter effect when the lower and upper microphone signals are summed.

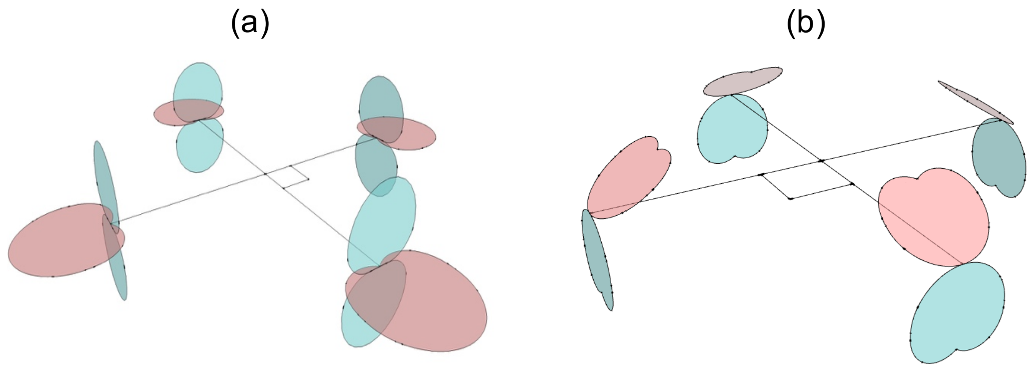

Fig. 9. Examples of the vertical extension of the quadraphonic ESMA for 3D sound capture (namely, ESMA-3D): (a) four vertical mid-side pairs of cardioid and fig-8 microphones (b) four vertical coincident pairs of cardioid microphones.

The first approach proposed here is to coincidentally arrange a vertically oriented figure-of-eight microphone with each of the main microphones of the ESMA. This is illustrated in Fig. 9 (a). Each of the vertical coincident pair is essentially a vertical mid-side pair. Therefore, it can be decoded into downward-facing and upward-facing virtual microphones, which are then routed to lower and upper loudspeakers in 3D sound reproduction, respectively, as described in [35]. When the microphone array is placed at the same height as the sound sources, the recommended loudspeaker arrangement is the so-called 'cube' format, which is commonly used for the 3D reproduction of an FOA recording (e.g., quadraphonic loudspeaker layers at -35° and 35° elevations). This will allow sound sources placed at the microphone array height to be presented as vertical phantom center images between the two loudspeaker layers, while sounds arriving from vertical directions would be localized vertically due to the ICLD cue.

In case of using the quadraphonic layer at the ear height augmented with another quadraphonic layer elevated at 30° to 45° [36], cardioid or supercardioid microphones facing directly upwards are recommended to capture the height information. Previous research suggests that to avoid the perceived position of a source image to be shifted upwards unintentionally in vertical stereophonic reproduction, the level of source sound captured by the height microphone needs to be at least 7–9 dB lower than that captured by the main microphone [37]. If the microphone array was raised at the same height as the sound source, with the main microphones being on-axis to the source, supercardioid microphones would be a better choice than cardioids for the height channels since they provide sufficient level attenuation for the source sound arriving from 90° (i.e., -10 dB). However, in case where the microphone array is raised higher than the sound source, which is common in classical music recording, cardioid microphones would also be suitable for the height channels since their theoretical polar response is smaller than -10 dB beyond 110° off-axis. In this case, it would be desired that the main microphones are angled on-axis towards the sources to ensure optimal localization and tonal quality, while the height microphones are angled directly upwards (e.g., Fig. 9 (b)). Note that this configuration makes the subtended angle between the main microphones of each stereophonic segment narrower than 90°, thus requiring a slight increase in microphone spacing to maintain the 90° SRA for each segment. For example, if the microphones of a quadraphonic ESMA are tilted downwards at -35.3°, the subtended angle for each microphone pair from the base point becomes 70.5° (consider the angle between the diagonals of a cube). In this case, based on the MARRS model [10], the correct spacing between the microphones to produce the 90° SRA is 54cm for cardioids. In case where a smaller array size is required, supercardioids could be used instead, which requires the microphone spacing of 40cm to achieve the SRA of 90°.

5 CONCLUSIONS

Listening experiments were conducted to evaluate the phantom image localization accuracies produced by different microphone spacings of the quadraphonic equal segment microphone array (ESMA) with cardioid microphones. The spacings of 24cm, 30cm and 50cm, which were based on different psychoacoustic models, as well as the 0cm spacing for the in-phase decoding of the first-order Ambisonics, were tested in both loudspeaker and binaural reproductions. The 50cm spacing was based on an ICTD and ICLD trade-off model optimized for the 90° loudspeaker base angle, whereas the 30cm and 24cm spacings were based on conventional models using data obtained for the 60° setup. The test stimuli were the recordings of an anechoic speech source located at 0° and 45° azimuth angles, made using the microphone arrays with the four different spacings as well as a dummy head. The listening tests measured the perceived positions of the phantom and real source images with the sound field rotated with 45° intervals, which was for simulating head-rotation or scene-rotation in virtual reality applications. The ITD and ILD produced in each experimental condition were also estimated.

From the results and discussions presented in this paper, the following conclusions are drawn:

(i) the 50cm spacing generally produces a more accurate and stable imaging than the other spacings tested, achieving the original design goal for an ESMA, which is the stereophonic recording angle of 90°. This also suggests that the conventional psychoacoustic models of ICTD and ILCD trade-off based on the 60° setup, which 30cm and 24cm microphone spacings were derived from, are not valid to be used for the quadraphonic setup. Therefore, the 50cm spacing is recommended to be used for a quadraphonic cardioid ESMA for recording 360° audio;

(ii) based on the binaural analysis of the stimuli, it seems that the 50cm spacing produces a better localization result due to a more effective ITD and ILD trade-off than the other spacings;

(iii) with the sound field rotation of the quadraphonic ESMA, a sound source placed at a central position tends to produce a less stable localization than that at a position closer to the microphones' on-axis directions (e.g. ±45°);

(iv) the binaural rendering of the ESMA recording produces more bimodal response distributions (e.g., front-back confusion) than the loudspeaker reproduction – this may be resolved by allowing head rotations in head-tracked VR scenarios.

Future work will examine the imaging accuracy of ESMA in a practical recording environment with a finer resolution of source angles. Furthermore, the octagonal ESMA and ESMA-3D designs described in Sections 4.5 and 4.6 will be evaluated. Investigations into the low-level spatial attributes of different 360° microphone arrays and their correlations with subjective preference and quality of experience in VR are currently ongoing. In addition, the influence of the acoustic characteristics of the recording venue on the perception of spatial attributes in 360° audio/visual recordings will be studied.

6 ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK, Grant Ref. EP/L019906/1. The author would like to thank the members of the Applied Psychoacoustics Lab (APL) at the University of Huddersfield who participated in the listening tests. He is also grateful to Tom Robotham for his drawings of the microphone arrays used in this paper.

7 REFERENCES

[1] F. Rumsey *Spatial Audio* (Oxford, Focal Press, 2001).

[2] E. Benjamin, R. Lee, and A. Heller, "Localization in horizontal-only Ambisonic system," presented at the *121st Convention of the Audio Engineering Society* (2006), convention paper 6967.

[3] E. Bates, M. Gorzel, L. Ferguson, H. O'dwyer and F. M. Boland, "Comparing Ambisonic Microphones: Part 1", presented at *the Audio Engineering Society Conference on Sound Field Control* (2016).

[4] M. Williams, "Microphone arrays for natural multiphony," presented at the *91st Convention of the Audio Engineering Society* (1991), convention paper 3157.

[5] M. Williams, "Migration of 5.0 multichannel microphone array design to higher order MMAD (6.0, 7.0&8.0) with or without the inter-format compatibility criteria," presented at the *124th Convention of the Audio Engineering Society* (2008), convention paper 7480.

[6] M. Williams and G. Le Du, "Microphone array analysis for multichannel sound recording," presented at the *107th Convention of the Audio Engineering Society* (1999), convention paper 4997.

[7] H. Lee, "Capturing and rendering 360° VR audio using cardioid microphones," presented at the *Audio Engineering Society Conference on Audio for Virtual and Augmented Reality* (2016).

[8] M. Williams, “Unified theory of microphone systems for stereophonic sound recording,” presented at the *82nd Convention of the Audio Engineering Society* (1987), convention paper 2466.

[9] H. Wittek and G. Theile, “The recording angle – based on localization curves,” presented at *the 112th Convention of the Audio Engineering Society* (2002 May), convention paper 5568.

[10] H. Lee, D. Johnson, and M. Mironovs, “An Interactive and Intelligent Tool for Microphone Array Design,” *presented at the 143rd Convention of the Audio Engineering Society* (2017 Oct), e-brief 390.

[11] G. Theile, “On the performance of two- channel and multi-channel stereophony,” presented at the *88th Convention of the Audio Engineering Society* (1990), convention paper 2932.

[12] H. Wittek, *Untersuchungen zur Richtungsabbildung mit L-C-R Hauptmikrofonen, Masters thesis*, Institut für Rundfunktechnik, 2000.

[13] G. Theile, “Multichannel natural recording based on psychoacoustic principles,” In *Proceedings of the Audio Engineering Society 19th International Conference*, pp. 201-229, 2001.

[14] H. Lee and F. Rumsey, “Level and time panning of phantom images for musical sources,” *J. Audio Eng. Soc.*, vol. 61 (12), pp. 753–767 (2013 Dec.).

[15] H. Lee, “Perceptually motivated amplitude panning (PMAP) for accurate phantom image localization,” presented at the *142nd Convention of the Audio Engineering Society* (2017), convention paper 9770.

[16] A. Farina, “Advancements in impulse response measurements by sine sweeps,” *presented at the 122nd Convention of the Audio Engineering Society* (2007 May), convention paper 7121.

[17] V. Hansen and G. Munch, “Making recordings for simulation tests in the Archimedes project,” *J. Audio Eng. Soc.*, vol. 39, pp. 768–774 (1991 Oct.).

[18] G. Kearney and T. Doyle, “An HRTF database for virtual loudspeaker rendering,” presented at the *139th Convention of the Audio Engineering Society* (2015), convention paper 9424.

[19] O. Kirkeby, P.A. Nelson, H Hamada, and F. Orduña Bustamante, “Fast deconvolution of multichannel systems using regularization,” *IEEE Trans. Speech and Audio Process.*, vol. 6 (2), pp.189–195 (March 1998). DOI: https://doi.org/10.1109/89.661479

[20] J. A. Hartigan and P. M. Hartigan, “The dip test of unimodality,” *Ann. Stat.*, vol. 13, pp. 70–84. (1985). DOI: https://doi.org/10.1214/aos/1176346577

[21] E. Benjamin, R. Lee, and A. Heller, “Is my decoder Ambisonics?,” presented at the *125th Convention of the Audio Engineering Society* (2008), convention paper 7553.

[22] G. Theile and G. Plenge, “Localization of lateral phantom images”, *J. Audio Eng. Soc.*, vol. 25, pp.196-200 (1977).

[23] G. Martin, W. Woszczyk, J. Corey and R. Quesnel, “Sound source localization in a five channel surround sound reproduction system’, presented at the *117th Convention the Audio Engineering Society* (1999), convention paper 4994.

[24] F. Wightman and D. Kistler, “Resolution of front–back ambiguity in spatial hearing by listener and source movement,” *J. Acoust. Soc. Am.,* vol. 105 (5), pp.2841-2853 (1999 May). DOI: https://doi.org/10.1121/1.426899

[25] P. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in *The Technology of Binaural Listening*, edited by Blauert, J. (Springer, Berlin, Heidelberg, 2013). DOI: https://doi.org/10.1007/978-3-642-37762-4

[26] L. R. Bernstein, and C. Trahiotis, “The normalized correlation: Accounting for binaural detection across center frequency,” *J. Acoust. Soc. Am.,* vol. 100 (5), pp.3774-3784 (1996).
DOI: https://doi.org/10.1121/1.417237

[27] V. Pulkki and M. Karjalainen, *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics* (Wiley, 2015).

[28] R. H. Whitworth and L. A. Jeffress, “Time versus intensity in the localization of tones,” *J. Acoust. Soc. Am.,* vol. 33, pp.925-929 (1961).
DOI: https://doi.org/10.1121/1.1908849

[29] J. Blauert, *Spatial hearing* (Cambridge, The MIT Press, 1997).

[30] D. Griesinger, “Stereo and Surround Panning in Practice,” *presented at the 112th Convention of the Audio Engineering Society* (2002 May), convention paper 5564.

[31] R. Wallis and H. Lee, “The effect of interchannel time difference on localization in vertical stereophony,” *J. Audio Eng. Soc.*, vol. 63 (10), pp.767-776 (2015 Oct.). DOI: https://doi.org/10.17743/jaes.2015.0069

[32] J. L. Barbour, “Elevation perception: phantom images in the vertical hemi-sphere,” In *Proceedings of the Audio Engineering Society 24th International Conference* (2003 June).

[33] M. Mironovs and H. Lee, “The influence of source spectrum and loudspeaker azimuth on vertical amplitude panning,” presented at the *142nd Convention of the Audio Engineering Society* (2017 May), convention paper 9782.

[34] H. Lee and C. Gribben, “Effect of vertical microphone layer spacing for a 3D microphone array” *J. Audio Eng. Soc.*, vol. 62, pp. 870–884 (2014 Dec.).
DOI: https://doi.org/10.17743/jaes.2014.0045

[35] P. Geluso, "Capturing height: the addition of Z microphones to stereo and surround microphone arrays," presented at the *132nd Convention of the Audio Engineering Society* (2012 April), convention paper 8595.

[36] ITU-R, *Recommendation ITU-R BS.2051-1: Advanced sound system for programme production* (2017).

[37] R. Wallis and H. Lee, "The reduction of vertical interchannel crosstalk: the analysis of localization thresholds for natural sound sources," *Appl. Sci. 7*, 278 (2017). DOI: https://doi.org/10.3390/app7030278

THE AUTHOR

Hyunkook Lee

Hyunkook Lee is a Senior Lecturer (i.e., Associate Professor) for music technology courses at the University of Huddersfield, UK, where he founded and leads the Applied Psychoacoustics Laboratory (APL). He is also a sound engineer with 20 years of experience in surround recording, mixing and live sound. Dr. Lee's recent research advanced understanding about the perceptual mechanisms of vertical stereophonic localisation and image spread as well as the phantom image elevation effect. This helped develop new 3D microphone array techniques, vertical mixing/upmixing techniques and a virtual 3D panning method. His ongoing research topics includes 3D sound perception, soundscape in virtual reality, virtual acoustics and objective sound quality metrics. From 2006 to 2010, Dr. Lee was a Senior Research Engineer in audio R&D at LG Electronics, South Korea, where he contributed to the standardizations of MPEG audio codecs and developed spatial audio algorithms for mobile devices. He received a B.Mus degree in music and sound recording (Tonmeister) from the University of Surrey, UK, in 2002, and obtained his Ph.D in sptial audio psychoacoustics from the Institute of Sound Recording (IoSR) at the same University in 2006. Hyunkook has been an active member of the AES since 2001 and received the AES fellowship award at the 145th convention in 2018.